06-13-00

A

06/12/00 JC849 U.S. PTO

PATENT APPLICATION TRANSMITTAL LETTER
(Small Entity)

Docket No.
4555-103 US

TO THE ASSISTANT COMMISSIONER FOR PATENTS

JC839 U.S. PTO 09/591746 06/12/00

Transmitted herewith for filing under 35 U.S.C. 111 and 37 C.F.R. 1.53 is the patent application of:

Yu, S. Z. et al.

For: **SYSTEM FOR WIRELESS PUSH AND PULL BASED SERVICES**

Enclosed are:

- ☒ Certificate of Mailing with Express Mail Mailing Label No. **EL512718279US**
- ☒ **5** sheets of drawings.
- ☐ A certified copy of a application.
- ☒ Declaration ☒ Signed. ☐ Unsigned.
- ☒ Power of Attorney
- ☐ Information Disclosure Statement
- ☐ Preliminary Amendment
- ☒ **1** Verified Statement(s) to Establish Small Entity Status Under 37 C.F.R. 1.9 and 1.27.
- ☒ Other: **Assignment**

CLAIMS AS FILED

For	#Filed	#Allowed	#Extra	Rate	Fee
Total Claims	29	- 20 =	9	x $9.00	$81.00
Indep. Claims	4	- 3 =	1	x $39.00	$39.00
Multiple Dependent Claims (check if applicable) ☐					$0.00
				BASIC FEE	$345.00
				TOTAL FILING FEE	$465.00

- ☒ A check in the amount of **$465.00** to cover the filing fee is enclosed.
- ☒ The Commissioner is hereby authorized to charge and credit Deposit Account No. **13-2165** as described below. A duplicate copy of this sheet is enclosed.
  - ☐ Charge the amount of as filing fee.
  - ☒ Credit any overpayment.
  - ☒ Charge any additional filing fees required under 37 C.F.R. 1.16 and 1.17.
  - ☐ Charge the issue fee set in 37 C.F.R. 1.18 at the mailing of the Notice of Allowance, pursuant to 37 C.F.R. 1.311(b).

Dated: **June 12, 2000**

Signature

Diane Dunn McKay
Reg. No. 34,586
Mathews, Collins, Shepherd & Gould, P.A.
100 Thanet Circle, Suite 306
Princeton, NJ 08540
(609) 924-8555 Telephone
(609) 924-3036 Facsimile

cc:

4555-103 US

EXPRESS MAIL CERTIFICATE

"Express Mail" Mailing Label Number: EL512718279US

Express Mail Corporate Account Number: X079384

Date of Deposit: June 12, 2000

Type of Documents:

1. Acknowledgment Post Card;
2. "Express Mail" Certificate;
3. Patent Application Transmittal Letter (Small Entity);
4. Copy of the Specification (13 pages), Claims (7 pages) & Abstract (1 page);
5. 5 sheets of Drawings;
6. Declaration and Power of Attorney (signed);
7. Assignment & Recordation Form;
8. Verified Statement Claiming Small Entity Status; and
9. Checks in the amounts of $465.00 and $40.00.

I hereby certify that this paper or fee is being deposited with the United States Postal Service "Express Mail Post Office to Addressee" service under 37 CFR 1.10 on the date indicated above and is addressed to the Assistant Commissioner of Patents, Washington, D.C. 20231; BOX: Patent Application.

Diane Dunn McKay
(Typed or printed name of person mailing paper or fee)

(Signature of person mailing paper or fee)

VERIFIED STATEMENT (DECLARATION) CLAIMING SMALL ENTITY STATUS (37 CFR 1.9(f) AND 1.27 (d)) - NONPROFIT ORGANIZATION	Docket No. 4555-103

Serial No.	Filing Date	Patent No.	Issue Date
Herewith	June 12, 2000	TBD	TBD

Applicant/ Patentee: Kobayashi, H. et al.

Invention: SYSTEM FOR WIRELESS PUSH AND PULL BASED SERVICES

I hereby declare that I am an official empowered to act on behalf of the nonprofit organization identified below:

NAME OF ORGANIZATION: PRINCETON UNIVERSITY

ADDRESS OF ORGANIZATION: Office of Patents & Licensing
5th Floor, New South Bldg.
P.O. Box 36
Princeton, NJ 08544-0036

TYPE OF NONPROFIT ORGANIZATION:

- ☒ University or other Institute of Higher Education
- ☐ Tax Exempt under Internal Revenue Service Code (26 U.S.C. 501(a) and 501(c)(3))
- ☐ Nonprofit Scientific or Educational under Statute of State of The United States of America
  Name of State: Citation of Statute:
- ☐ Would Qualify as Tax Exempt under Internal Revenue Service Code (26 U.S.C. 501(a) and 501(c)(3)) if Located in The United States of America
- ☐ Would Qualify as Nonprofit Scientific or Educational under Statute of State of The United States of America if Located in The United States of America
  Name of State: Citation of Statute:

I hereby declare that the above-identified nonprofit organization qualifies as a nonprofit organization as defined in 37 C.F.R. 1.9(e) for purposes of paying reduced fees to the United States Patent and Trademark Office regarding the invention described in:

- ☒ the specification to be filed herewith.
- ☐ the application identified above.
- ☐ the patent identified above.

I hereby declare that rights under contract or law have been conveyed to and remain with the nonprofit organization with regard to the above identified invention.

If the rights held by the above-identified nonprofit organization are not exclusive, each individual, concern or organization having rights to the invention is listed on the next page and no rights to the invention are held by any person, other than the inventor, who could not qualify as an independent inventor under 37 CFR 1.9(c) or by any concern which would not qualify as a small business concern under 37 CFR 1.9(d) or a nonprofit organization under 37 CFR 1.9(e).

Each person, concern or organization to which I have assigned, granted, conveyed, or licensed or am under an obligation under contract or law to assign, grant, convey, or license any rights in the invention is listed below:

☒ no such person, concern or organization exists.
☐ each such person, concern or organization is listed below.

FULL NAME _____
ADDRESS _____
☐ Individual ☐ Small Business Concern ☐ Nonprofit Organization

FULL NAME _____
ADDRESS _____
☐ Individual ☐ Small Business Concern ☐ Nonprofit Organization

FULL NAME _____
ADDRESS _____
☐ Individual ☐ Small Business Concern ☐ Nonprofit Organization

FULL NAME _____
ADDRESS _____
☐ Individual ☐ Small Business Concern ☐ Nonprofit Organization

Separate verified statements are required from each named person, concern or organization having rights to the invention averring to their status as small entities. (37 CFR 1.27)

I acknowledge the duty to file, in this application or patent, notification of any change in status resulting in loss of entitlement to small entity status prior to paying, or at the time of paying, the earliest of the issue fee or any maintenance fee due after the date on which status as a small entity is no longer appropriate. (37 CFR 1.28(b))

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application, any patent issuing thereon, or any patent to which this verified statement is directed.

NAME OF PERSON SIGNING: Allen J. Sinisgalli
TITLE IN ORGANIZATION:
ADDRESS OF PERSON SIGNING: 5th Floor, New South Bldg., P.O. Box 36, Princeton, NJ 08544-0036

SIGNATURE: Allen J. Sinisgalli per DATE: 5/30/00

SYSTEM FOR WIRELESS PUSH AND PULL BASED SERVICES

Background of the Invention

1. Technical Field

The invention relates to a proxy gateway for providing improved push and pull based services from a content provider on the Internet to a mobile user on a wireless network.

2. Description of Related Art

The Internet is a global network formed by the cooperative interconnection of computing networks. The Worldwide Web (WWW or Web) is a collection of files or "Web pages" of text, graphics and other media which are connected by hyperlinks to other Web pages which physically reside on the Internet. In a transaction on the WWW, a Web client typically requests information from a Web server. The requested information is transmitted from the Web server to the Web client over the Internet. Dramatically increasing expansion of Internet services using the WWW has led to increased Web traffic.

A conventional technique to reduce Web traffic and speed up Web access is to store copies of documents in a cache. U.S. Patent No. 5,873,100 describes an Internet browser which includes an embedded cache for user controlled document retention. The cache stores a plurality of documents. At least one of the documents stored in the cache is designated as a keep document. If the storage limit of the cache is exceeded, the cache deletes the oldest document not designated as a keep document.

Web servers and Web client use hypertext transfer protocol (HTTP) / 1.1 which includes cache control features. See R. Fielding et al., "Hypertext Transport Protocol HTTP / 1.1" Network Working Group RFC, May 1996, URL: ftp://ftp.isi.edu/in-notes/rfc2068.txt. The original Web server assigns expiration times to responses generated for Web client requests. An expiration judgment is performed in the cache when a cached entry is requested by a client. If the cached entry has not expired, the cache sends the entry to the client; otherwise, it sends a conditional request to the Web

server. A validation check is performed at the Web server to validate if the cached entry is still useable. If the cached entry is useable, the Web server sends a validator to the Web client; otherwise, it sends an updated response.

In certain systems, there exists very little local memory. In these systems, caching and prefetching is preferably performed at a proxy server intermediate between the Web server and the Web client. Prefetching is a technique in which additional items are fetched when a request is made for a particular item. U.S. Patent No. 5,925,100 describes a system having a dumb client environment in which a smart server determines when to send a prefetched object to the user. The prefetched objects are determined based on an object based prefetch primitive present in the client's executing application.

Other conventional prefetch schemes are based on predicting at a given time the likelihood that a given document will be accessed in the near future. Prefetch schemes have been described in which a prediction module computes the access probability that a file will be requested in the near future. Each file whose access probability exceeds a server's prefetch threshold is prefetched. See Z. Jiang and L. Kleinrock, "An Adaptive Network Prefetch Scheme," *IEEE J. on Selec. Areas in Common.*, vol. 16, no. 3, April 1998, pp. 358-368, and Z. Jiang and L. Kleinrock, "Web Prefetching in a Client Environment," *IEEE Personal Communications*, vol. 5, no. 5, Oct. 1998, pp. 25-34.

In addition, prefetch schemes have been described which are based on popularity based prefetching. See E. P. Markatos, "Main Memory Caching of Web Documents," *Computer Networks and ISDN Systems*, vol. 28, issues 7-11, pp. 893-906, 1996. Main menu caching of frequently requested documents is performed on the Web server. Similarly, U.S. Patent No. 5,991,306 describes a pull based intelligent caching system for delivering data over the Internet. Content of frequently requested documents is downloaded from the content provider and cached at a local service provider. The content is cached prior to a peak time when subscribers are likely to request the content. A pattern recognizer detects behavior patterns based on subscriber requests to determine which content the subscribers are most likely to request and when. When content is finally requested, the data is streamed continuously for rendering at the subscriber computer.

Another prefetching scheme type is based on interactive prefetching in which prefetching is determined by the interaction between the client and the server. For example, an interactive prefetching scheme has been proposed in which the system gathers references by passing hypertext markup language (HTML) in the referenced page and collecting referenced pages with each request. See K. Chinen and S. Yamaguchi, "An Interactive Prefetching Proxy Server for Improvement of WWW Latency, " *INET'97*, Kuala Lumpur, Malaysia, 1997.

Wireless systems and mobile users are typically limited to small bandwidth and small memory. A wireless application protocol (WAP) has been developed to promote industry-wide specifications for technology useful in developing applications and services, such as Internet services, that operate over wireless communication networks, as described in Wireless Architecture Protocol Specification, Wireless Application Protocol Forum, Ltd., Version 30, April 1998 and WAP Push Architectural Overview, Version 08, November 1999. WAP framework defines pull technology as a transaction initiated by the client to pull information from a server. For example, the World Wide Web is an example of pull technology in which a user enters a universal resource locator (URL) which is sent to the server and the server sends a Web page to the user. WAP framework defines push technology as a transmission to the client without previous action by the client. Accordingly, the server pushes information to the client without an explicit request from the client. It is desirable to provide a system for wireless push and pull based Internet services which expeditiously allows users to gain access to desired Web information.

Summary of the Invention

The present invention relates to a method and system for providing Web content from pull and push based services running on Web content providers to mobile users. A proxy gateway connects the mobile users to the Web content providers. A prefetching module is used at the proxy gateway to optimize performance of the pull services by reducing average access latency. The prefetch module prioritizes pull content to be stored in a cache at the proxy gateway. The average access latency can be reduced by using at least one factor related to the frequency of access to the pull content, the update

cycle of the pull content determined by the Web content providers and the response delay for fetching pull content from the content provider to the proxy gateway. Pull content, such as documents, having the greatest average access latency are sorted and a predetermined number of the documents are prefetched into the cache. Push services are optimized by iteratively estimating a state of each of the mobile users to determine relevant push content to be forward to the mobile user. The estimate of the state of each mobile user can be determined from tracking information of the mobile user and geo-location measurement and behavior observation data.

The invention will be more fully described by reference to the following drawings.

Brief Description of the Drawings

Fig. 1 is a schematic diagram of a system for providing push and pull based services to mobile users.

Fig. 2 is a flow diagram of an implementation of a prefetch module used in a proxy gateway of the system.

Fig. 3 is an illustration of a caching model.

Fig. 4 is a schematic diagram for providing push services of the system.

Fig. 5 is a flow diagram of an implementation of the mobile state prediction.

Detailed Description

Reference will now be made in greater detail to a preferred embodiment of the invention, an example of which is illustrated in the accompanying drawings. Wherever possible, the same reference numerals will be used throughout the drawings and the description to refer to the same or like parts.

Fig. 1 illustrates a schematic diagram of a system for providing push and pull based services to mobile users 10. Mobile users 12a-12n are connected by mobile network 11 to proxy gateway 13. For example, mobile network 11 is a wireless network. Web server 14 is connected by network 15 to proxy gateway 13. For example, network 15 can be a wired network, such as the Internet.

Each of mobile users 12a-12n can interact with pull service 16 and push service 20 running on content provider 14, such as a Web server. As an example of pull service 16, mobile user 12a generates a pull request 17 which is transmitted over network 11 to gateway 13. Pull request 17 is transmitted via proxy gateway 13 to pull service 16. Mobile user 12a receives pull response 18 generated by pull service 16 at content provider 14 via proxy gateway 13. Pull response 18 includes pull content 19 stored or generated by content provider 14. As an example of push service 20, push message 21 is sent by push service 20 to mobile user 12b. Push message 21 includes push content 22 stored or generated at content provider 14. Push message 21 is sent by content provider 14 to proxy gateway 13. Push message 21 is forwarded by proxy gateway 13 to mobile user 12b.

Pull service 16 and push service 20 interact with cache 23, as described below. Pull response 18 generated by content provider 14 can be prefetched and cached in cache 23. Thereafter, when pull request 17 identifies a pull response 18 stored in cache 23, pull response 18 stored in cache 23 can be forwarded to mobile user 12a, or any other mobile user 12b-12n, without contacting content provider 14. Prefetch module 24 reduces access latency of pull service 16 using document access log data 25 to determine prefetch control information 26 for prefetching a predetermined number of pull responses 18 into cache 23. Unprefetched pull response 18 is cached by using a conventional caching mechanism, such as HTTP/1.1 caching. Mobile state estimation pedication module 27 uses mobile tracking data 28 and geo-location measurement and behavior observation data 29 to determine push control information 30a for controlling caching of push messages 21 in cache 23 and for determining timing for sending relevant push content to the mobile user 30b depending on the state of mobile user 12b. It will be appreciated that multiple pull services 16 and push services 20 can be provided at content provider 14 or a plurality of content providers 14 and accessed by multiple mobile users 12a-12n. Pull content 19 and push content 22 can comprise, for example, documents including Hypertext Markup Language (HTML) files, Java files, and embedded data including images, video files and audio files.

Fig. 2 illustrates a flow diagram for performing prefetch module 24. Prefetch module 24 optimizes performance of pull service 16 by reducing average access latency

using at least one factor related to the frequency of access to pull content 19, the update cycle of pull content 19, and the response delay for fetching pull content 19 from Web server 14 to cache 23. Access latency is defined as the time between when pull request 17 is transmitted from a mobile user 12a-12n to the time when pull response 18 is received at a mobile user 12a-12n. Web server 14 generates content header 31 which is forwarded as a header of pull response 18. For example, content header 31 can include fields as defined in HTTP/1.1 such as: Expires time, relating to the expiration time of pull content 19; Last-Modified time, relating to the time pull content 19 was last modified at Web server 14; and Content-Length, relating to the size of pull content 19. Proxy gateway 13 receives content header 31 and stores fields of content header 31 in document access log data 25. Proxy gateway 13 also receives pull request 17 and stores fields related to pull request 17. Fields related to cache 23 are also stored in document access log data 25. For example, fields stored in document access log data 25 can include fields such as: request_time, related to the local time when proxy gateway 13 made pull request 17 to content provider 14; and response_time, related to the local time when cache 23 received pull response 18 from content provider 14.

In block 32, variables related to pull service 16 are extracted from document access log data 25. For example, the extraction block 35 can calculate the parameters associated with potential candidates that may be included into the list of the prefetched documents. The document that has only one access record (i.e., no revisit) in the log or has not more than one update cycle during the statistic period is not treated as a potential candidate. Let R_n be the average rate of access to pull content 19. Pull content 19 is referred to as document n. R_n can be determined at proxy gateway 13 based on log data of cache 23 and content header 31 of pull responses 18 from content providers 14 and stored in document access log data 25. Let s_n be the size of document n. For example, s_n can be determined from the Content-Length field forwarded by content provider 14 and stored in document access log data 25.

Let ΔT_n be the average response delay imposed by network 15 which can be measured by the time from when pull request 17 from a mobile user 12a-12n is reformatted and forwarded by proxy gateway 13 to content provider 14 to when pull content 19 is fetched from or validated by content provider 14 to cache 23, i.e. which can

be measured by the difference between the logged request_time and response_time. $T_{s,n}$ is the average time delay imposed by network 11 and network 15, which is the time from when, pull request 17 is generated from mobile user 12a where pull request 17 is reformatted and is forwarded by proxy gateway 13 to content provider 14 and pull content 19 is fetched from or validated by content provider 14 as pull response 18 to proxy gateway 13 to when response 18 is received by mobile user 12a. $T_{c,n}$ is the average time delay imposed by network 11, which is the time between the generation of pull request 17 from mobile user 12a and the reception of pull response 18 by mobile user 12a when a valid copy of pull content 19, document n, is found in cache 23, including transmission time of pull response 18, roundtrip time from cache 23 to mobile user 12a and cache 23 processing time. Approximately, $\Delta T_n = T_{s,n} - T_{c,n}$. Let μ_n be the update cycle of pull content 19, document n, which is the average length of time between two successive expiration times or two successive modifications of pull content 19, document n.

The caching model of HTTP/1.1 is illustrated in Fig. 3 in which t_k represents the Last_Modified time and "o" represents the Expires time of pull content 19, document n generated by Web server 14. Cache 23 is accessed by a group of mobile users 12a-12n and N is the total number of documents n residing in various Web servers 14 with n = 1, ... N. Pull request 17a generated by one of mobile users 12a-12n and forwarded to cache 23 in cycle, μ_n, cannot be satisfied by cache 23 which is denoted by missing 1. Cache 23 fetches a copy of pull content 19 for pull request 17a from Web server 14. Consequent pull requests 17b and 17c generated by one of mobile users 12a-12n in cycle, μ_n, are satisfied by cache 23 which is denoted by hit 1 and hit 2. Thereafter, in a second cycle, μ_n, pull request 17d generated by one of mobile users 12a-12n and forwarded to cache 23 in cycle, μ_n, cannot be satisfied by cache 23 which is denoted by missing 2. Cache 23 fetches a fresh copy of pull content 19 for pull request 17d from Web server 14. Consequent pull requests 17e-g generated by one of mobile users 12a-12n in cycle, μ_n, are satisfied by cache 23 which are denoted respectively by hits 3-5. In a third cycle, μ_n, pull request 17h generated by one of mobile users 12a-12n and forwarded to cache 23 in cycle, μ_n, cannot be satisfied by cache 23 which is denoted by missing 3. Thereafter, pull request 17i generated by one of mobile users 12a-12n arrives at cache 23 between the

expiration time and end of cycle, μ_n. In this case, cache 23 validates pull content 19 at Web server 14, represented by validate 1, before using pull content 19 stored in cache 23. The distribution of interarrival time of pull requests 17a-17i to pull content 19 represented by document n is represented by $f_n(t)$ which can be an exponential distribution. Since Web server 14 typically specifies the expires time based on its schedule to the end of cycle, μ_n, the interval between the Expires time and the end of the cycle can have a stochastic or deterministic distribution.

In block 34 of Fig. 2, the access probability of access to document n, represented by γ_n is determined by:

$$\gamma_n = R_n / R \qquad (1)$$

wherein R is the total rate of access traffic on network 15 from gateway 13, which is the sum of R_n for n=1, 2, ..., N.

In block 36, the average hit rate for document n, represented by h_n, is determined by:

$$h_n = 1 - \frac{g_n}{R_n \mu_n}, \quad n=1, 2, \ldots, N, \qquad (2)$$

in which:

g_n is the probability that there is at least one request to document n during a given update cycle, μ_n, given by:

$$g_n = 1 - e^{-R_n \mu_n}, \quad n = 1, 2, \ldots, N, \qquad (3)$$

and

$R_n\mu_n$ is the expected number of accesses to document n in an update cycle of document n.

In block 38, the wired network access latency, represented by η_n, imposed by network 15 when the request is the first one for document n in the update cycle μ_n or Expires time for document n has been exceeded, as shown in Fig. 3, is computed from:

$$\eta_n = \gamma_n (1-h_n) \Delta T_n, \quad n = 1, \ldots, N, \qquad (4)$$

The values at η_n, n = 1, ..., N are sorted in descending order with document 1 having the greatest average latency imposed by network 15 labeled as η_1, and document N having the least average latency labeled as η_N and relabeled as: $\eta_1 \geq \eta_2 \geq \eta_3 \geq \ldots \geq \eta_N$.

In block 40, the total number of documents n to be prefetched to cache 23, represented by r, is determined by considering at least one factor. Examples of factors include: spare capacity of cache 23 that can be utilized by the prefetching after providing sufficient capacity for conventional caching, ΔC; spare transmission bandwidth, ΔB, of network 15; and desired improvement of hit probability, ΔH. The total number of documents to be prefetched represented by r satisfies all of the following constraints:

The constraint of spare cache capacity, ΔC, is given by

$$\sum_{n=1}^{r} s_n (1-h_n) \leq \Delta C; \qquad (5)$$

where $\Delta C \approx C - \sum_{n=1}^{N} s_n h_n$, C is given capacity of cache 23 and $\sum_{n=1}^{N} s_n h_n$ is the capacity required for conventional caching such as described in the caching model of HTTP/1.1.

The constraint of spare transmission bandwidth, ΔB, on network 15 is given by:

$$\sum_{n=1}^{r} (1-g_n) \frac{s_n}{\mu_n} \leq \Delta B; \qquad (6)$$

where$\Delta B \approx B - \sum_{n-1}^{N} g_n \frac{s_n}{\mu_n}$, B is given bandwidth and $\sum_{n-1}^{N} g_n \frac{s_n}{\mu_n}$ is the bandwidth required for conventional caching.

The constraint of desired minimum improvement of hit probability, ΔH, is given by:

$$\sum_{n=1}^{r} \gamma_n (1-h_n) \geq \Delta H \qquad (7)$$

where $\Delta H \approx H - \sum_{n=1}^{N} \gamma_n h_n$, H is given hit probability, and $\sum_{n=1}^{N} \gamma_n h_n$ is the total average hit probability for conventional caching.

In block 41, the total number of documents to be prefetched to cache 23, r, that correspond to the r largest are selected and relabeled as $\eta_1, \ldots, \eta_r$. In block 42, the documents determined in block 41 are prefetched into cache 23 at proxy gateway 13 as

soon as the documents expire at cache 23, which means the prefetched document n in cache 23 is updated with average cycle, μ_n, as shown in block 43. Thereafter, blocks 32-43 are repeated with a given period, such as several hours, several days, or several times the expected update cycle, $\bar{\mu}$, is represented by:

$$\bar{\mu} = \sum_{h=1}^{N} \gamma_n \mu_n \qquad (8')$$

Accordingly, average latency for pull service 16 is derived as:

$$L = \sum_{n-1}^{N} \gamma_n [h_n T_{c,n} + (1-h_n) T_{s,n}] - \sum_{n-1}^{r} \eta_n \qquad (8)$$

In Eq. (8), the first term is the latency of a conventional cache scheme and is independent of the prefetch scheme used and the second term $\sum_{n-1}^{r} \eta_n$ is the latency reduction as a result of prefetch scheme of the present invention. The latency reduction is determined by the number r and the selection of r prefetched documents. Thus, performance of blocks 32-43 minimizes the average latency L.

Fig. 4 illustrates a schematic diagram for providing push services of the system. Movement and behavior of the mobile user 50 is measured by geo-location measurement and behavior observation block 29. An example of movement and behavior of the mobile user data 50 is represented in Table 1 illustrating examples of different states a mobile user 12a-12n can occupy.

Table 1 Mobile States

State	Description	Position and Behavior	Time	Speed	Direction	Mean dwell time in the state
State 0	Inactive (power off or out off location-dependent services).	--	--	--	--	d_0
State 1	Walking on a street.	X_1	t_1	≈1m/s	along one direction	d_1
State 2	In a shopping mall.	X_2	open hours	≈0		d_2
State 3	Drive on a highway.	X_3	t_3	≈30m/s	along one direction	d_3
State M-1						d_{MH}

State 0 is an inactive state in which mobile user 12a-12n can occupy when powering off its mobile terminal or is out of location-dependent services of wireless network 11. State 1 to State M-1 are active states which mobile user 12a-12n can occupy while being actively involved with network 11 and interacting with pull service 16 and push service 20. Each state is determined by several parameters such as: position and behavior of mobile user 12a-12n, at time t represented by X_t, time of determining state, represented by t; speed of mobile user 12a-12n; direction of movement of mobile user 12a-12n; and mean dwell time in state m, represented by d_m wherein M is the total number of distinct states defined for mobile user 12a-12n, and m=0, 1, ..., M-1. Y_t represents the results of geo-location position measurement and behavior observations of mobile user 12a-12n at time t. It is observed that Y_t is the observed or measured value of the position and behavior of mobile user 12a-12n and is in general different from X_t which is the true, but unknown, position and behavior because of geo-location and estimation error. Geo-location position and behavior data 29 can be estimated with conventional methods as described in J.H. Reed, K.J. Krizman, B.D. Woerner and T.S. Rappaport, "An Overview of the Challenges and Progress in Meeting the E-911 Requirement for Location Service", IEEE Communications Magazine, Vol. 36, No. 4, April 1998, pp. 30-37, hereby incorporated by reference into this application.

S_t is the state of mobile user 12a-12n at time t wherein $S_t \in \{0, 1, ..., M-1\}$.

Mobile user 12a-12n can transit from one state to another. The transit mobility of one of mobile users 12a-12n, mobile user 12, can be represented by an M-state Markov chain with transition probability matrix (TPM):

$$P=[p_{mn}]_{MXM} \qquad (9)$$

where the p_{mn} is the probability of transition from state m to state n, and m, n=0, 1, ..., M-1.

Tracking data 28 represents sequence data of mobile user 12 over time. Tracking data 28 includes Y_1^t and $\{\alpha_\tau(m), \tau=1, ..., t, m=1, ..., m-1\}$

S_1^t represents the state sequence of mobile user 12 from time 1 to t. X_1^t represents the corresponding position and behavior sequence of mobile user 12 from time 1 to t.

Y_1^t represents the corresponding geo-location and observation sequence of mobile user 12 from time 1 to t. Mobile state prediction module 27 can generate push control in information 30 for controlling caching of push content 22 and for determining timing for sending push content 30b in cache 23 to mobile user 12 based on determined states.

Fig. 5 is a flow diagram of an implementation of mobile state prediction module 27. Mobile state prediction module 27 estimates the posteriori probability of the states of mobile user 12a-12n for a given geo-location and observation sequence, Y_1^t.

In block 60, the initial state of mobile user 12a-12n is defined upon registration of mobile user 12a-12n in system 10. Let $\alpha_t(m)$ represent a forward variable for state sequence estimation which is a probability that the State at time t is m and the corresponding geo-location and observation sequence is Y_1^t.

The forward variables for state sequence estimation for mobile user 12 in the initial time are:

$$\alpha_0(0)=1, \qquad \alpha_0(m)=0, \qquad \text{for } m=1, 2, \ldots, M-1.$$

In block 62, a measured or observed value of a current geo-location position and behavior of mobile user 12, Y_t, is determined for determining geo-location measurement and behavior observation data 29.

In block 63, the probability $\Pr\{Y_t \mid X\}$ that the geo-location measured result is Y_t when mobile user 12 position and behavior is X is predetermined by the geolocation and observation error distribution.

In blocks 65 the values required for iteration are stored in a database which values can be mobile tracking data 28. In block 64, state sequence estimation variable $\alpha_t(m)$ for all m=1, 2 ... M-1 is determined, which is stored into the database for the next recursive computation.

The following analysis can be used for performing block 64. At time t, tracking data is represented by: $Y_1^t = (Y_1, Y_2, \ldots, Y_{t-1}, Y_t)$ where Y_t is the current measured data of the position and the behavior of mobile user 12 from the geo-location measurement and the collected information of proxy gateway 13. $\alpha_t(m)$ is computed for all m=1, 2, ..., M-1, by iterations from 1 through t from the following:

$$\alpha_t(m) = \sum_{m'=0}^{M-1} \Pr\{S_{t-1} = m'; S_t = m; Y_1^t\}, \qquad (10)$$

$$= \sum_{m'=0}^{M-1} \Pr\{S_{t-1} = m';\ Y_1^{t-1}\}\ \Pr\{S_t = m;\ Y_t | S_{t-1} = m'\}, \qquad (11)$$

$$= \sum_{m'=0}^{M-1} \alpha_{t-1}(m')\ p_{m'm} \sum_{x} \Pr\{x | m\} \Pr\{Y_t | x\} \qquad (12)$$

wherein $p_{m'm}$ is the state transition probability of mobile user 12, $\Pr\{x|m\}$ is the probability that the mobile user locates at position and behavior as x when it is in state m at time t, for example the output probability of the Markov source, and $\Pr\{Y_t|x\}$ is the probability that the geo-location measured result is Y_t when the mobile's position and behavior is x at time t.

In block 66, state z is determined by:

$$z = \arg \max_m \{\Pr\{S_t = m | Y_1^t\} | m = 1,2,...,M-1\}, \qquad (13)$$

$$= \arg \max_m \{\frac{\Pr\{S_t = m; Y_1^t\}}{\Pr\{Y_1^t\}} | m = 1,2,...,M-1\}, \qquad (14)$$

$$= \arg \max_m \{\alpha_t(m) | m = 1,2,...,M-1\}. \qquad (15)$$

Accordingly, blocks 60, 63, 64 and 66 determine mobile state prediction module 27.

In block 68, state z related push content 22 is pushed to mobile user 12a-12n as relevant push content to the mobile user 30b. In block 69, mobile user 12 may accept or reject push content 22 and this behavior is observed by block 62. The mobile user's current behavior and position is forwarded to block 62 and block 62-69 can be repeated.

It is to be understood that the above-described embodiments are illustrative of only a few of the many possible specific embodiments which can represent applications of the principles of the invention. Numerous and varied other arrangements can be readily devised in accordance with these principles by those skilled in the art without departing from the spirit and scope of the invention.

We Claim:

1. A method for providing at least one pull service and at least one push service to a plurality of mobile users comprising the steps of:

reducing access latency for said at least one pull service running on at least one Web server by prefetching documents into a cache of at least one proxy gateway by using at least one factor relating to a frequency of access of said plurality of mobile users to said pull content of said pull service, an update cycle of said pull content and response delay for fetching said pull content from said at least one Web server to at least one proxy gateway, said at least one proxy gateway connected between said mobile user and said Web server; and

iteratively estimating a state of each of said plurality of mobile users for determining push content to be forwarded to said mobile user by said at least one push service running on said at least one Web server.

2. The method of claim 1 wherein said pull content is plurality of documents and said step of reducing access latency comprises the step of selecting a predetermined number of documents to be prefetched into said cache of said proxy gateway, wherein said predetermined number of documents have the greatest reduction in said access latency.

3. The method of claim 1 wherein said step of reducing access latency uses said factor of said frequency of access wherein frequently accessed documents are prioritized for being stored in a cache of a proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service.

4. The method of claim 1 wherein said step of reducing access latency uses said factor of said update cycle wherein said pull documents having a shorter update cycle are prioritized for being stored into a cache of a proxy gateway said proxy gateway being connected between said mobile user and said pull service and push service.

5. The method of claim 1 wherein said access latency uses said factor of said response delay wherein said pull documents having a longer response delay are prioritized for being stored in a cache of a proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service.

6. The method of claim 1 wherein said step of reducing access latency comprises the step of selecting a predetermined number of documents to be prefetched into cache of a proxy gateway, and said step of selecting a predetermined number of documents uses said factors of: said frequency of access, said update cycle and said response delay, wherein said frequently accessed pull documents having a shorter update cycle and a longer response delay are prioritized for being prefetched in said cache of said proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service.

7. The method of claim 1 wherein said step of iteratively estimating a state of said mobile user is determined from tracking data of said plurality of mobile users and geo-location measurement and behavior observation data.

8. The method of claim 7 further comprising the step of:

caching mobility and behavior-related push content into a cache of said proxy gateway connected between said plurality of mobile users and said at least one Web server.

9. The method of claim 8 wherein each said state of one of said mobile users is determined from at least one of the following factors: location of said one of said plurality of mobile users, direction of said one of said plurality of mobile users, speed of said one of said plurality of mobile users, and behavior of said one of plurality of mobile users.

10. A system for providing at least one pull service and at least one push service to a plurality of mobile users comprising:

means for reducing access latency for said at least one pull service running on at least one Web server by prefetching documents into a cache of a proxy gateway, said proxy gateway being connected between said mobile user and said pull service and push service by using at least one factor relating to a frequency of access of said plurality of mobile users to said pull content of said pull service, an update cycle of said pull content and response delay for fetching said pull content from said at least one Web server to said at least one proxy gateway said at least one proxy gateway connected between said mobile user and said Web server; and

means for iteratively estimating a state of each of said plurality of mobile users for determining push content to be forwarded to said mobile user by said at least one push service running on said at least one Web server.

11. The system of claim 10 wherein said pull content is a plurality of documents and said means for reducing access latency comprises the step of:

selecting a predetermined number of documents to be prefetched into a cache of a proxy gateway; and

selecting a predetermined number of documents uses said factors of: said frequency of access, said update cycle and said response delay, wherein said pull documents having a higher frequency of access, a shorter update cycle and a longer response delay are prioritized for being prefetched in said cache of said proxy gateway.

12. The system of claim 10 wherein said means for iteratively estimating a state of said mobile user uses tracking data of said plurality of mobile users and geo-location measurement and behavior observation data.

13. The system of claim 12 further comprising:

means for controlling of caching mobility and behavior-related push content into a cache of said proxy gateway connected between said plurality of mobile users and said at least one Web server.

14. The system of claim 13 wherein each said state of one of said mobile users is determined from at least one of the following factors: location of said one of said plurality of mobile users, direction of said one of said plurality of mobile users, speed of said one of said plurality of mobile users, and behavior of said one of plurality of mobile users.

15. In a system comprising a proxy gateway connected by a first network to a plurality of mobile users and by a second network to at least one Web server, said proxy gateway comprising a cache for storing pull content received from said at least one Web server of a pull service, a method comprising the steps of:

storing data that is indicative of a request for said pull content from at least one of said plurality of mobile users and data indicative of interactions between said cache and said Web server;

determining access probability of access to said pull content from said stored data;

determining an average hit rate for said pull content from said stored data;

determining said average response delay for said pull content from said stored data;

determining average wired network access latency for said pull content from said access probability, said average hit rate and said average response delay;

storing said pull content in said cache based on said determined average wired network access latency when there is no said pull content in said cache or said pull content has expired,

wherein said pull content having a greater average wired network access latency is prioritized for being stored in said cache.

16. The method of claim 15 wherein said pull content is a plurality of n documents, n=1, 2 ... N, wherein N is the total number of documents, and said stored data comprises:

an average rate of access to document n, R_n; a size of said document n, s_n; an average time delay imposed by said second network, ΔT_n; and an update cycle of said document n, μ_n.

17. The method of claim 16 wherein said access probability is determined by:

$$\gamma_n = R_n / R$$

wherein R is the total rate of access traffic on said second network.

18. The method of claim 17 wherein said average hit rate for document n, h_n is determined by:

$$h_n = 1 - \frac{g_n}{R_n \mu_n}, \quad n=1, 2, \ldots, N,$$

in which:

g_n is the probability that there is at least one request to document n during a given update cycle, μ_n, given by:

$$g_n = 1 - e^{-R_n \mu_n}, \quad n = 1, 2, \ldots, N, \qquad (3)$$

and

$R_n\mu_n$ is the expected number of accesses to document n in an update cycle of document n.

19. The method of claim 18 wherein average wired-network-access latency when there is no said pull content in said cache or said pull content has expired is determined from

$$\eta_n = \gamma_n (1-h_n) \Delta T_n, \quad n = 1, \ldots, N,$$

20. The method of claim 19 wherein said pull content is prioritized by the steps of:

sorting said plurality of N documents in descending order with the document having the greatest average wired network access latency when there is no said pull content in said cache or said pull content has expired labeled as η_1, and the document having the least average wired-network-access latency when there is no said Web content in said cache or said Web content has expired labeled as η_N; and

determining a number of documents to be stored in said cache, r, by considering at least one constraint selected from the group consisting of spare cache capacity, spare transmission bandwidth on said second network and desired hit probability.

21. The method of claim 20 wherein said constraint of said spare cache capacity, ΔC, is given by:

$$\sum_{n=1}^{r} s_n (1-h_n) \leq \Delta C,$$

wherein $\Delta C \approx C - \sum_{n=1}^{N} s_n h_n$, C is given capacity of the cache, s_n is the size of the document n and h_n is the average hit rate for document n.

22. The method of claim 20 wherein said constraint of said spare transmission bandwidth, ΔB, is given by:

$$\sum_{n=1}^{r} (1-g_n) \frac{S_n}{\mu_n} \leq \Delta B,$$

wherein $\Delta B \approx B - \sum_{n=1}^{N} g_n \frac{S_n}{\mu_n}$, B is given bandwidth, g_n is the probability that there is at least one request to document n during a given update cycle μ_n and s_n is the size of the document.

23. The method of claim 20 wherein said constraint of said desired minimum hit probability, ΔH, is given by:

$$\sum_{n=1}^{r} \gamma_n \ (1\text{-}h_n) \geq \Delta H$$

wherein $\Delta H \approx H - \sum_{n=1}^{N} \gamma_n \ h_n \ H$ is given hit probability, γ_n is an access to document n and h_n is an average hit rate for document n.

24. The method of claim 16 further comprising the step of:

updating said stored pull content in said cache based on said update cycle of document n, μ_n.

25. In a system comprising a proxy gateway connected by a first network to a plurality of mobile users and by a second network to at least one Web server, a method comprising the steps of:

measuring each of said mobile users current geo-location position and behavior;

computing a first probability that said measured current geo-location position and behavior is an actual position and behavior of each of said mobile users;

determining a state sequence estimation variable for each of said mobile users by iteration over time from a second probability that each of said mobile users transit in a geo-location and behavior sequence;

determining a current state for each of said mobile users from said state sequence estimation; and

pushing push content related to said current state to each of said mobile users.

26. The method of claim 25 wherein said first probability is given by

$$P_r \{Y_t \mid X\}$$

wherein Y_t is said measured current geo-location position and behavior and X is said actual position and behavior.

27. The method of claim 26 wherein said state sequence estimation variable is determined by

$$\alpha_t(m) = \sum_{m'=0}^{M-1} \alpha_{t-1}(m') \ p_{m'm} \sum_{x} \Pr\{x \mid m\} \Pr\{Y_t \mid x\}$$

wherein $p_{m'm}$ is the state transition probability of one of said plurality of mobile users, $\Pr\{x \mid m\}$ is the probability that said one of said plurality of mobile users locates at position and behavior as x when it is in state m at time t, and $\Pr\{Y_t \mid x\}$ is the probability that said

measured geo-location is Y_t when said one of said plurality of mobile users position and behavior is x at time t.

28. The method of claim 27 wherein said current state is determined by

$z = \arg \max_m \{\alpha_t(m) \,|m = 1,2,...,M-1\}$.

29. The method of claim 28 wherein said proxy gateway comprising a cache for storing push content received from said at least one Web server and said push content is stored in said cache based on said current state.

Abstract of the Disclosure

The present invention relates to a method and system for providing Web content from pull and push based services running on Web content providers to mobile users. A proxy gateway connects the mobile users to the Web content providers. A prefetching module is used at the proxy gateway to optimize performance of the pull services by reducing average access latency. The average access latency can be reduced by using at least three factors: one related to the frequency of access to the pull content; second, the update cycle of the pull content determined by the Web content providers; and third, the response delay for fetching pull content from the content provider to the proxy gateway. Pull content, such as documents, having the greatest average access latency are sorted and a predetermined number of the documents are prefetched into the cache. Push services are optimized by iteratively estimating a state of each of the mobile users to determine relevant push content to be forward to the mobile user.

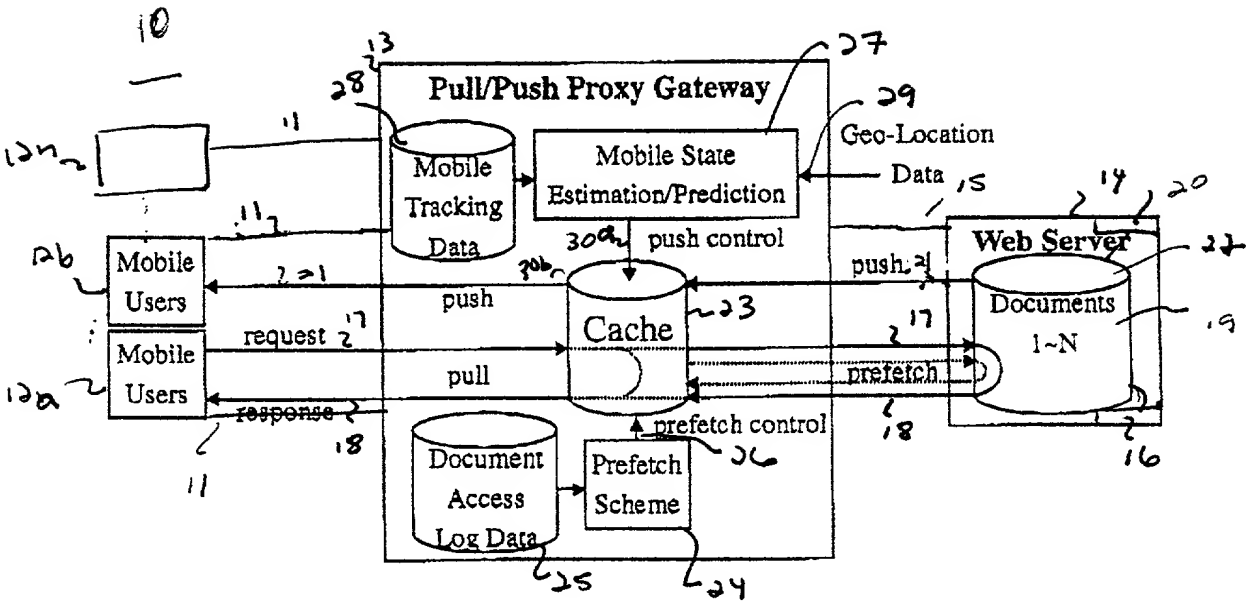
10
13
28
Pull/Push Proxy Gateway
27
29
Geo-Location
Data
15
Mobile
Tracking
Data
Mobile State
Estimation/Prediction
push control
30a
30b
Web Server
14
20
22
Documents
1~N
19
Mobile
Users
12b
Mobile
Users
12a
push
21
request
17
pull
response
18
Cache
23
prefetch
17
18
16
prefetch control
26
Document
Access
Log Data
25
Prefetch
Scheme
24

Figure 1

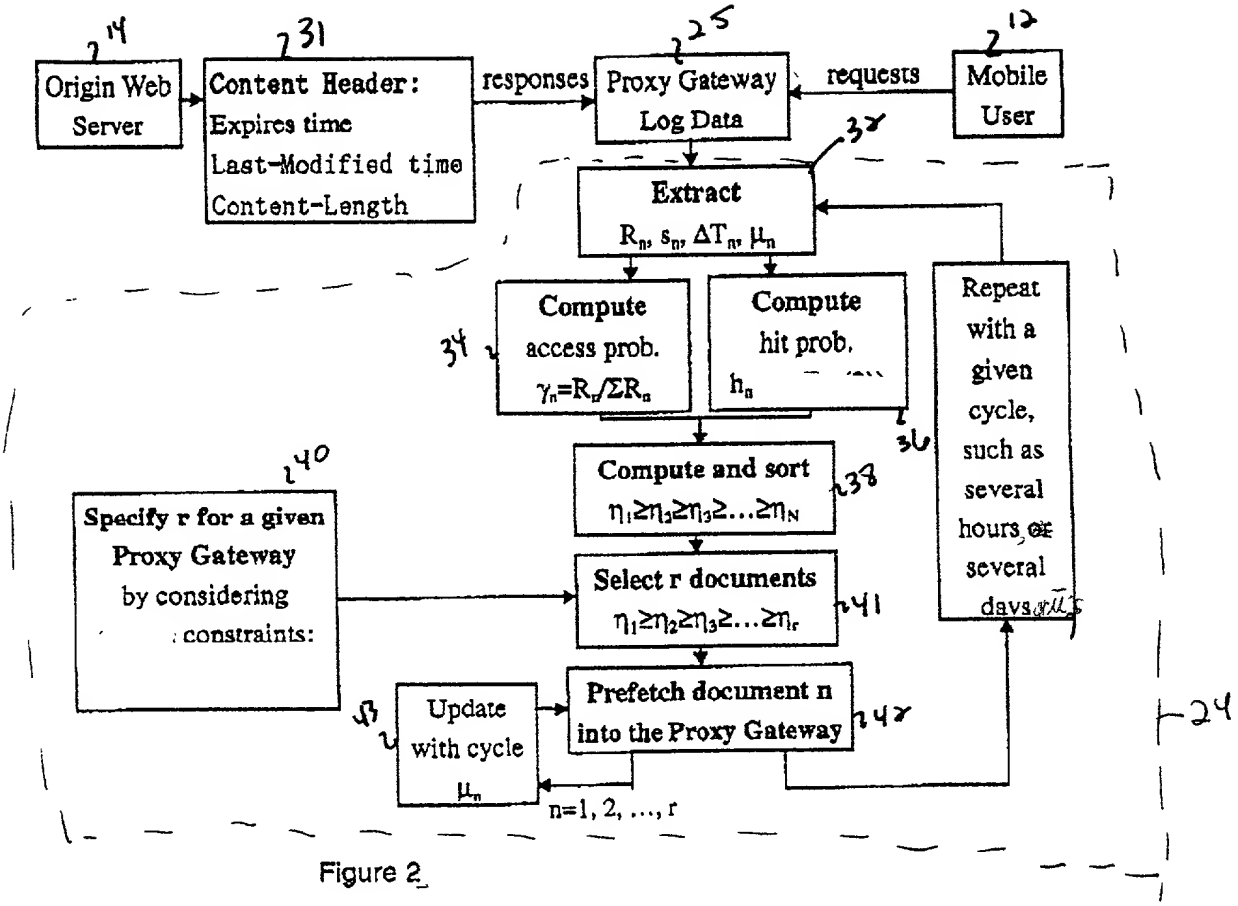
14
31
25
12
Origin Web Server
Content Header:
Expires time
Last-Modified time
Content-Length
responses
Proxy Gateway Log Data
requests
Mobile User
37
Extract
R_n, s_n, ΔT_n, μ_n
Compute access prob.
$\gamma_n=R_n/\Sigma R_n$
Compute hit prob.
h_n
34
36
Repeat with a given cycle, such as several hours or several days
Compute and sort
$\eta_1\geq\eta_2\geq\eta_3\geq\ldots\geq\eta_N$
38
40
Specify r for a given Proxy Gateway by considering constraints:
Select r documents
$\eta_1\geq\eta_2\geq\eta_3\geq\ldots\geq\eta_r$
41
Update with cycle μ_n
Prefetch document n into the Proxy Gateway
42
43
n=1, 2, ..., r
24

Figure 2

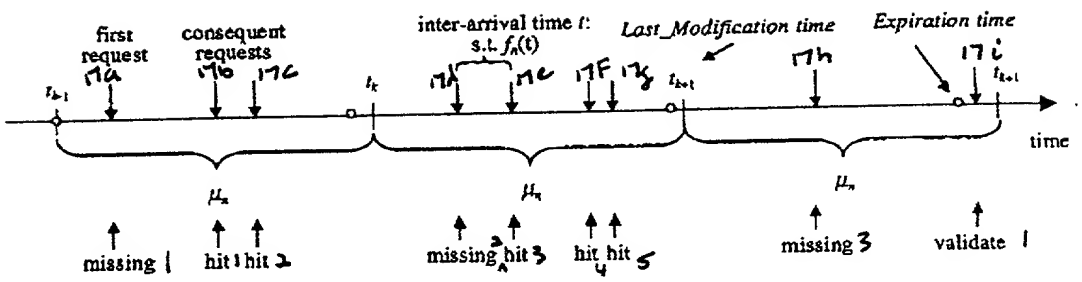
first request
consequent requests
inter-arrival time t: s.t. $f_n(t)$
Last_Modification time
Expiration time
time
missing
hit
hit
missing
hit
hit
hit
missing
validate

Fig. 3

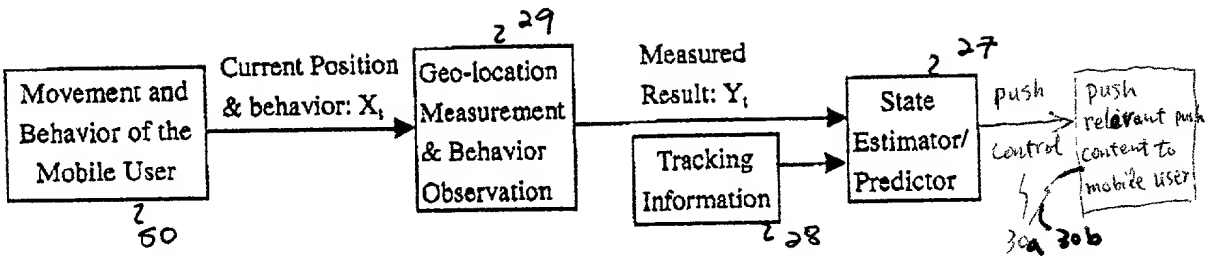
Movement and Behavior of the Mobile User
50
Current Position & behavior: X_t
Geo-location Measurement & Behavior Observation
29
Measured Result: Y_t
Tracking Information
28
State Estimator/ Predictor
27
push
control
push relevant push content to mobile user
30a
30b

Fig. 4

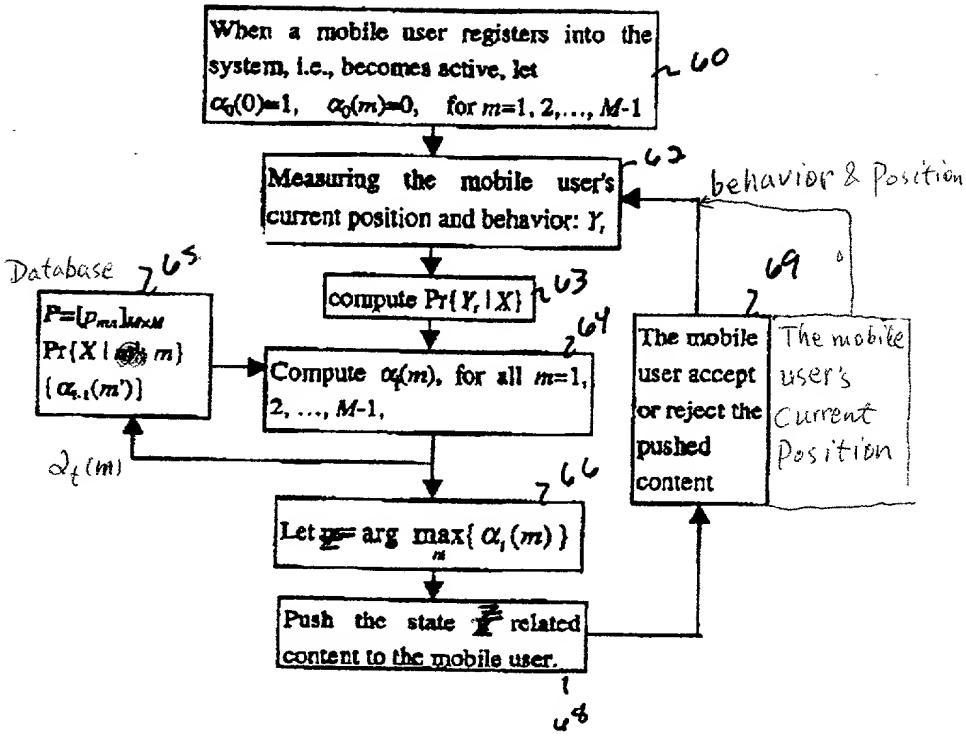
When a mobile user registers into the system, i.e., becomes active, let $\alpha_0(0)=1$, $\alpha_0(m)=0$, for $m=1, 2, \ldots, M-1$
60
Measuring the mobile user's current position and behavior: Y_t
62
behavior & Position
Database
65
$P=[p_{mn}]_{M\times M}$
$\Pr\{X \mid m\}$
$\{\alpha_{t-1}(m')\}$
compute $\Pr\{Y_t \mid X\}$
63
Compute $\alpha_t(m)$, for all $m=1, 2, \ldots, M-1$,
64
$\alpha_t(m)$
Let $x = \arg\max_m \{\alpha_t(m)\}$
66
Push the state x related content to the mobile user.
68
69
The mobile user accept or reject the pushed content
The mobile user's Current Position

FIG. 5

Docket No.
4555-103 US

Declaration and Power of Attorney For Patent Application

English Language Declaration

As a below named inventor, I hereby declare that:

My residence, post office address and citizenship are as stated below next to my name,

I believe I am the original, first and sole inventor (if only one name is listed below) or an original, first and joint inventor (if plural names are listed below) of the subject matter which is claimed and for which a patent is sought on the invention entitled
SYSTEM FOR WIRELESS PUSH AND PULL BASED SERVICES

the specification of which

(check one)

☒ is attached hereto.

☐ was filed on _____ as United States Application No. or PCT International Application Number _____

and was amended on _____
(if applicable)

I hereby state that I have reviewed and understand the contents of the above identified specification, including the claims, as amended by any amendment referred to above.

I acknowledge the duty to disclose to the United States Patent and Trademark Office all information known to me to be material to patentability as defined in Title 37, Code of Federal Regulations, Section 1.56.

I hereby claim foreign priority benefits under Title 35, United States Code, Section 119(a)-(d) or Section 365(b) of any foreign application(s) for patent or inventor's certificate, or Section 365(a) of any PCT International application which designated at least one country other than the United States, listed below and have also identified below, by checking the box, any foreign application for patent or inventor's certificate or PCT International application having a filing date before that of the application on which priority is claimed.

Prior Foreign Application(s)			Priority Not Claimed
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	☐
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	☐
_____ (Number)	_____ (Country)	_____ (Day/Month/Year Filed)	☐

I hereby claim the benefit under 35 U.S.C. Section 119(e) of any United States provisional application(s) listed below:

(Application Serial No.)	(Filing Date)

I hereby claim the benefit under 35 U. S. C. Section 120 of any United States application(s), or Section 365(c) of any PCT International application designating the United States, listed below and, insofar as the subject matter of each of the claims of this application is not disclosed in the prior United States or PCT International application in the manner provided by the first paragraph of 35 U.S.C. Section 112, I acknowledge the duty to disclose to the United States Patent and Trademark Office all information known to me to be material to patentability as defined in Title 37, C. F. R., Section 1.56 which became available between the filing date of the prior application and the national or PCT International filing date of this application:

(Application Serial No.)	(Filing Date)	(Status) (patented, pending, abandoned)

I hereby declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

POWER OF ATTORNEY: As a named inventor, I hereby appoint the following attorney(s) and/or agent(s) to prosecute this application and transact all business in the Patent and Trademark Office connected therewith. *(list name and registration number)*

Bruce M. Collins, Reg. No. 20,066
Ronald Gould, Reg. No. 28,299
Diane Dunn McKay, Reg. No. 34,586
Glen E. Books, Reg. No. 24,950
Anastasia P. Winslow, Reg. No. 40,875
Mary S. Kakefuda, Reg. No. 39,245

For the firm:
Mathews, Collins, Shepherd & Gould, P.A.
100 Thanet Circle, Suite 306
Princeton, NJ 08540
(609) 924-8555 Telephone
(609) 924-3036 Facsimile

Send Correspondence to: **Diane Dunn McKay**
Mathews, Collins, Shepherd & Gould, P.A.
100 Thanet Circle, Suite 306
Princeton, NJ 08540

Direct Telephone Calls to: *(name and telephone number)*
Diane Dunn McKay (609) 924-8555

Full name of sole or first inventor
Shun Zheng Yu

Sole or first inventor's signature Shunzheng Yu Date June 5, 2000

Residence 30-01 Fox Run Dr., Plainsboro, NJ 08536, USA

Citizenship Chinese

Post Office Address

Full name of second inventor, if any
Hisashi Kobayashi

Second inventor's signature Hisashi Kobayashi Date June 5, 2000

Residence 21 Russell Road, Princeton, NJ 08540, USA

Citizenship Japanese

Post Office Address